Copy of Prior Foreign Application (#2003108434)

СПОСОБ ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ ИЗОБРАЖЕНИЯ

МАШИНОЧИТАЕМОЙ ФОРМЫ НЕФИКСИРОВАННОГО ФОРМАТА.

(A METHOD OF IMAGE PRE-ANALYZING OF A MACHINE-READABLE FORM OF NON-FIXED LAYOUT.)

US application is a authentic translation of Russian Prior document

Class G06K009/34

СПОСОБ ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ ИЗОБРАЖЕНИЯ МАШИНОЧИТАЕМОЙ ФОРМЫ НЕФИКСИРОВАННОГО ФОРМАТА.

Преобразование информации с бумажных носителей в электронную форму осуществляется системами оптического распознавания символов - системами OCR (Optical Character Recognition). В настоящее время системы распознавания достаточно устойчиво *распознают тексты высокого и среднего качества, напечатанные* любым стандартизированным шрифтом. Это позволяет использовать их для преобразования больших массивов текстовой информации, иногда без последующего контроля человеком. Однако посимвольное преобразование является недостаточным для полноценного перевода информации в электронный вид. Поэтому основной задачей, решаемой существующими системами OCR является идентификация логической структуры документов, т.е. явное выделение реквизитов документа - заголовка, полей ввода данных и пр.

Основная сложность решения задачи - недостаточная формализация существующих бумажных документов. Даже если перечень реквизитов документа строго фиксирован, например, как на платежном поручении, бухгалтерских балансах, налоговой декларации и других документах, расположение областей изображения, содержащих реквизиты, специфицируется лишь приблизительно с некоторым допуском.

Кроме того, точное расположение полей трудно обеспечить при массовом тиражировании. Таким образом, идентификация логической

структуры документов в большинстве случаев сводится к задаче распознавания изображения документа с нестрого фиксированным расположением реквизитов.

Пространственно печатный документ формируется из структурных элементов: текстовых областей, различных графических объектов и разделительных линий. Некоторые текстовые области относятся к заполняемым реквизитам. Другие текстовые области, графические объекты и разделительные линии относятся к элементам формы.

Свойства пространственной структуры могут быть описаны с использованием пространственных и параметрических характеристик структурных элементов:

- ограничений на абсолютное расположение структурного элемента;
- ограничений на относительное расположение структурного элемента;
- ограничений на изменение геометрических размеров структурного элемента;
- присутствие структурного элемента может быть факультативным.

Абсолютное расположение структурных элементов характерно для анкет и стандартных бланков. При этом фиксировано и их относительное взаимное расположение. При этом остается задача точного определения положения всей страницы с учетом сдвигов, поворотов и помех, возникающих при сканировании. В остальных случаях ограничения на абсолютное расположение могут быть

использованы для отсечения заведомо неправильных вариантов идентификации структурных элементов.

Относительные ограничения могут быть разделены на две группы:

- качественные;
- количественные.

Качественные ограничения фиксируют общий характер отношения, например, что один элемент находится выше или ниже другого. Количественные ограничения задают координаты области, в которой расположен элемент, относительно другого элемента. Как правило, количественными ограничениями задают также малые отклонения во взаимном расположении. В остальных случаях применяют качественные отношения.

В некоторых случаях стандарты формы не являются достаточно строгими и не фиксируют взаимное расположение реквизитов, что допускает возможность структурных вариаций.

Весьма распространенным случаем структурных вариаций можно считать отсутствие того или иного структурного элемента на изображении. Часто это вызвано необязательностью его заполнения. Изображение элемента может также деградировать при сканировании, так что его идентификация становится невозможной. Примером таких элементов являются разделительные линии, текстовые элементы бланка, напечатанные мелким шрифтом. Тем не менее, использовать такие элементы при анализе необходимо, так как они позволяют дополнительно уточнить положение других элементов.

Изобретение относится к области оптического распознавания символов машиночитаемых форм в полях ввода данных и, в частности, к способам подготовки изображения и/или шаблона к проведению операций распознавания текста из растрового изображения в случае, когда положение полей ввода данных не строго фиксировано.

Широко известен способ предварительной обработки при распознавании изображения документа, при котором растровое изображение разбивают на области, содержащие текст, и области содержащие нетекстовые объекты.

Разработчики существующих систем идут на ограничение и упрощение модели документа для того, чтобы иметь возможность воспользоваться существующими формальными методами структурного распознавания образов.

Патент США №5864629 (January 26, 1999, Wustmann) решает задачу выделения логической структуры документа для частного случая, когда в интересующих областях имеются уникальные символы, не встречающиеся в других областях документа. В патенте задача проиллюстрирована на примере областей, содержащих символ «$».

Способ недостаточно универсален и приспособлен для решения единственной очень частной задачи.

В патенте США №6507671 (January 14, 2003, Kagan, и др.) описан способ выделения на изображении машиночитаемой формы заполняемой части и удаления остального содержимого формы.

Способ описан на примере обработки стандартной машиночитаемой формы, имеющей фиксированное расположение полей.

Недостатком способа является низкое быстродействие, в связи с тем, что этап выделения логической структуры совмещен с этапом распознавания текста в полях формы.

Известны способы, использующие фиксированное расположение полей. Они часто применяются в системе распознавания форм, *однако имеют существенный недостаток - ограничение области* применения специально спроектированными машиночитаемыми формами.

Например, в патенте США №5822454 (October 13, 1998, Rangarajan) раскрыт способ распознавания формы документа единственного типа - СЧЕТ (Invoice). Документ имеет строго фиксированную форму как по составу полей, так и по их свойствам, расположению и длине.

Способ недостаточно универсален и не приспособлен для обработки форм, допускающих отклонения в составе и пространственных характеристиках полей ввода данных.

Технический результат изобретения состоит в расширении возможностей обработки машиночитаемых форм, в особенности не имеющих строго фиксированного расположения полей.

Известные способы не позволяют эффективно выделять поля ввода на изображении формы.

Указанные недостатки, значительно ограничивают возможности использования известных способов для выделения полей ввода на изображении формы.

Известные способы непригодны для достижения заявленного технического результата.

Указанный технический результат достигается тем, что на изображении машиночитаемой формы предварительно назначают графические или текстовые элементы, которые затем используют как реперы для пространственной привязки полей ввода данных. Информацию о пространственных характеристиках реперов и об относительном расположении полей ввода данных помещают в описание шаблона формы. После сканирования изображение машиночитаемой формы разбивают на области содержащие изображения связных областей, линий, и др. объектов. Определяют положение реперов на изображении формы. Определяют положение полей ввода данных относительно реперов. В случае множественного результата идентификации поля ввода выбирают наиболее близко соответствующий по пространственным и параметрическим характеристикам.

Сущность предложения проиллюстрирована на фиг. 1 и фиг. 2.

На фиг. 1 показан документ с назначенными легко различимыми объектами формы в качестве реперов.

На фиг. 2 показана пространственная привязка поля ввода данных к назначенному объекту формы - реперу.

Способ состоит в следующем.

На изображении машиночитаемой формы предварительно назначают графические или текстовые элементы (1), которые затем используют как реперы для пространственной привязки полей ввода данных (2).

Назначенные элементы (1) (реперы) должны быть хорошо идентифицируемы после ввода формы с помощью сканера. Это могут быть графические и/или текстовые элементы.

Информацию о пространственных характеристиках реперов и об относительном расположении полей ввода данных (2) и их допусках для объекта каждого типа помещают в описание его шаблона.

Кроме того, в описание шаблона помещают параметрическую информацию о полях ввода, например, длину символьного поля, интервал допустимых значений, соотношения с другими полями и др.

Графическое изображение машиночитаемой формы после предварительной обработки по исправлению наклона, сдвига, очистки от помех, разбивают на области содержащие изображения полей ввода, связных областей, линий, символов и др. объектов.

Определяют положение реперов на изображении формы. Если репер является текстовой областью, дополнительно проводят распознавание ее содержания.

Определяют положение полей ввода данных относительно реперов. Если с учетом допусков схожие и/или однотипные поля пересекаются, происходит их частичное или полное наложение, выбирают наиболее близко соответствующее по пространственным и параметрическим характеристикам. Если этого недостаточно, привлекают дополнительную параметрическую информацию о полях ввода.

Процесс идентификации выполняют путем выдвижения и проверки гипотез о принадлежности поля.

Репер может быть простым - состоящим из одного объекта - и составным - состоящим более, чем из одного объекта.

Кроме того, возможно задание репера в виде альтернативы. То есть репер задают как один из вариантов - графический объект типа «прямоугольник» либо текстовый объект, содержащий текст «Плательщик», либо графический объект типа «разделительная линия» и т.д.

Возможно также объединение вышеуказанных способов задания репера, например, группа 1, либо группа 2, либо группа 3 и т.д.

Возможно также в качестве репера использовать ранее найденное поле ввода данных.

Оценка достоверности вариантов распознавания, т.е. гипотез, осуществляется по следующему алгоритму распознавания.

Для того, чтобы сравнивать и комбинировать оценки различных структурных элементов, необходимо привести их к единой шкале. Принято интерпретировать оценку достоверности как оценку условной вероятности:

$Q_t \sim p(t|I)$, где t - структурный элемент, а I - изображение.

На основании вероятностной интерпретации оценки достоверности принимают:

$$p(N|I) = p(N|t_1,\ldots,t_n)p(t_1|I)\ldots p(t_n|I),$$

где $p(t_i|I)$ - вероятность i-го подэлемента составного элемента N,

$p(N|t_1;\ldots,t_n)$ - условная вероятность того, что данный набор подэлементов образует составной элемент N.

Оценку достоверности распознавания составного элемента вычисляют по следующему правилу:

$Q = Q_R Q_1 \ldots Q_n$,

где $Q_i \sim p(t_i|I)$ - оценка достоверности i-го подэлемента составного элемента N.

$Q_R \sim p(N|t_1,\ldots,t_n)$ - оценка отношений

Величина $p(N|t_1,\ldots,t_n)$ соответствует отношениям Q_R, которые ограничивают подэлементы в составном элементе. Отношения должны моделировать функцию условной плотности вероятности $p(N|t_1,\ldots,t_n)$. Для решения данной задачи используют инструментарий нечеткой логики в вероятностной интерпретации. Логические операции вычисляют оценку по следующим формулам:

операция '$\wedge$' по формуле: $A \wedge B \rightarrow a * b$;

операция '$\vee$' по формуле: $A \vee B \rightarrow a + b - a * b$;

операция '$\neg$' по формуле: $\neg A \rightarrow 1 - a$.

Для формирования функции условной плотности вероятности используют кусочно-линейную аппроксимацию.

ФОРМУЛА

1. Способ предварительной обработки изображения машиночитаемой формы с нефиксированным размещением полей, характеризующийся

наличием изображения заполненной формы,

наличием по крайней мере одного шаблона формы с описанием пространственных и параметрических свойств объектов формы,

выполнением следующих этапов:

- этапа устранения сдвига, наклона изображения и компенсации искажений,
- этапа разбиения изображения на области,
- этапа выделения областей, содержащих поля ввода данных;

отличающийся выполнением следующих операций:

- предварительного назначения на форме по крайней мере одного объекта формы в качестве репера для пространственной привязки по крайней мере одного поля ввода данных,
- описания пространственных характеристик по крайней мере одного указанного репера в описании шаблона формы,
- идентификации на изображении формы по крайней мере одного репера,
- определения положения по крайней мере одного поля ввода относительно по крайней мере одного указанного репера.

2. Способ по п. 1, отличающийся тем, что репер представляет собой текстовую область.

3. Способ по п. 2, отличающийся тем, что дополнительно проводят распознавание текстовой области, используемой в качестве репера.

4. Способ по п. 1, отличающийся тем, что при множественном результате поиска идентификацию поля проводят путем выдвижения и проверки гипотез и оценкой качества соответствия описанию в шаблоне.

5. Способ по п. 4, отличающийся тем, что привлекают дополнительную параметрическую информацию о поле ввода.

6. Способ по п. 1, отличающийся тем, что этап определения положения по крайней мере одного поля ввода относительно по крайней мере одного указанного репера в свою очередь включает по крайней мере следующие этапы:

- выбора поля для поиска из описания шаблона,
- выбора из описания шаблона характеристик по крайней мере одного репера для пространственной привязки искомого поля,
- поиска по крайней мере одного указанного репера на изображении формы,
- поиска указанного поля на изображении формы относительно по крайней мере одного репера с учетом пространственных и параметрических характеристик поля, описанных в шаблоне,
- идентификации указанного поля из нескольких, в случае множественного результата поиска.

7. Способ по п. 1, отличающийся тем, что поле ввода может быть репером для поиска других полей.

8. Способ по п. 1, отличающийся тем, что этап идентификации поля из нескольких, удовлетворяющих набору пространственных и параметрических характеристик выполняют полностью или частично с привлечением оператора.

9. Способ по п. 1, отличающийся тем, что пространственное положение репера не фиксировано.

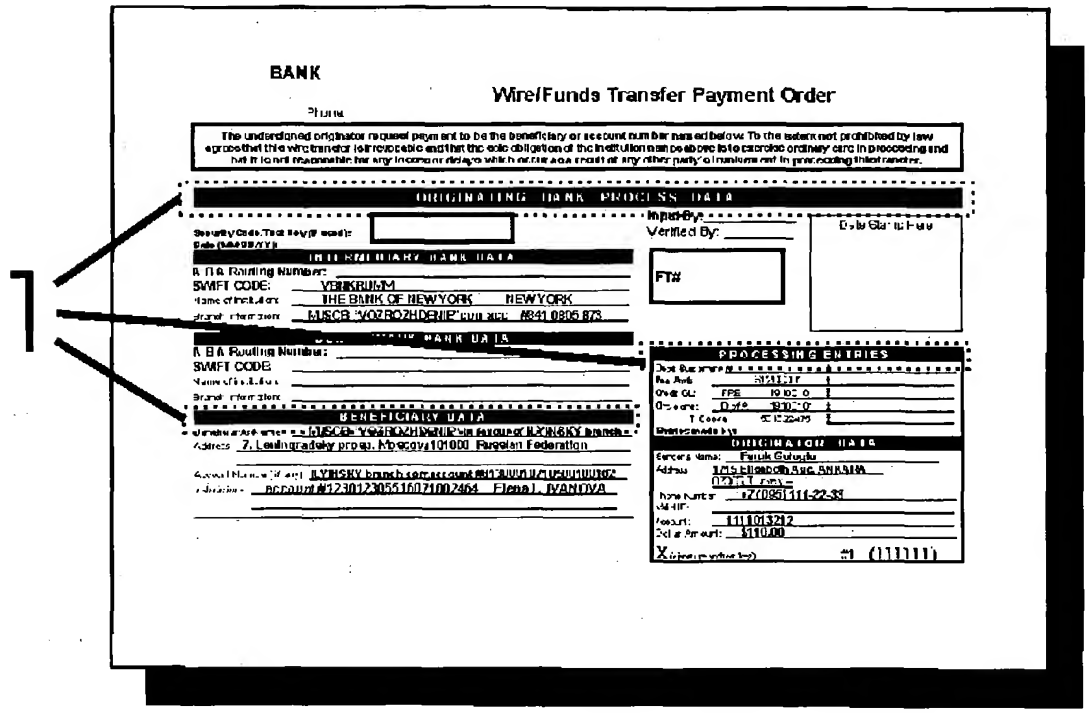
10. Способ по п. 1, отличающийся тем, что один репер используют для пространственной привязки более одного поля.

11. Способ по п. 1, отличающийся тем, что пространственную привязку поля осуществляют к более, чем одному реперу.

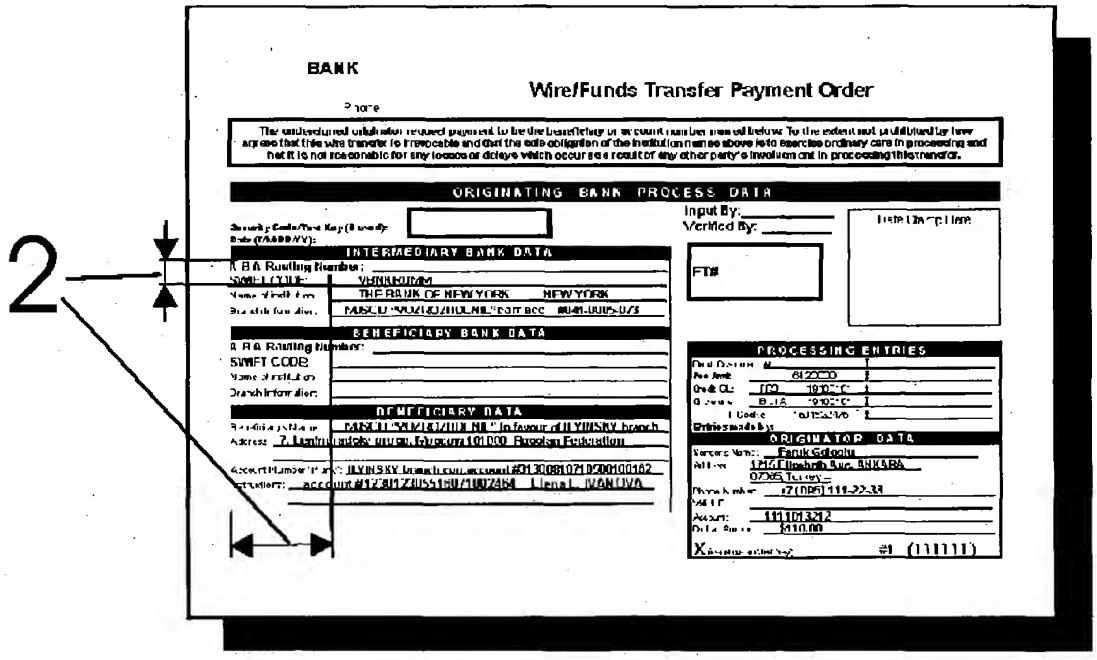
12. Способ по п. 1, отличающийся тем, что репер включает более одного объекта.

13. Способ по п. 1, отличающийся тем, что репер описывают в виде альтернативы.

СПОСОБ ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ ИЗОБРАЖЕНИЯ МАШИНОЧИТАЕМОЙ ФОРМЫ НЕФИКСИРОВАННОГО ФОРМАТА.



ФИГ. 1



ФИГ. 2